Patent Retrieval in Chemistry based on Semantically Tagged Named Entities

Harsha Gurulingappa[1,2], Bernd Müller[1,2], Roman Klinger[1], Heinz-Theodor Mevissen[1], Martin Hofmann-Apitius[1,2], Juliane Fluck[1], and Christoph M. Friedrich[1]

[1]Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)
Schloss Birlinghoven 53754 Sankt Augustin, Germany
[2]Bonn-Aachen International Centre for Information Technology (B-IT)
Dahlmannstraße 2, 53113 Bonn, Germany

Abstract

This paper reports on the work that has been conducted by Fraunhofer SCAI for TREC Chemistry (TREC-CHEM) track 2009. The team of Fraunhofer SCAI participated in two tasks, namely *Technology Survey* and *Prior Art Search*. The core of the framework is an index of 1.2 million chemical patents provided as a data set by TREC. For the technology survey, three runs were submitted based on semantic dictionaries and noun phrases. For the prior art search task, several fields were introduced into the index that contained normalized noun phrases, biomedical as well as chemical entities. Altogether, 36 runs were submitted for this task that were based on automatic querying with tokens, noun phrases and entities along with different search strategies.

1 Introduction

Text processing in chemistry is more formidable in comparison to other fields due to the presence of different possibilities to represent chemical name mentions such as trivial names, IUPAC[1] [8], brand names, InChI[2], and SMILES [7]. For example, the drug name "Aspirin" is reported to have 25 synonyms and 95 brand names in DrugBank[3]. In order to address this challenge, TREC provides a workbench for large scale evaluation and comparison of different techniques for text retrieval in chemistry. TREC-CHEM addresses this challenge in terms of two independent tasks. The first task, namely the technology survey, consists of 18 expert-defined natural language expressions of the information needed and the task is to retrieve a set of documents from a predefined collection that can best answer the questions. The second task, namely prior art search, consists of 1000 test patents and the task is to retrieve sets of documents invalidating each test patent.

Considering the ambiguity inherent to the chemistry-based literature, our approach focused on tagging the chemical and biomedical named entities in the documents. Tagging the entities and mapping them to standard database entries normalizes different forms of the same entity to one standard form. This helps to overcome the problems associated with multiple synonyms, acronyms and morphological variants in text. Moreover, document retrieval based on semantically tagged entities has demonstrated variable success in the past [4, 10, 11]. A precondition for such an approach is the availability of comprehensive and domain specific dictionaries as well as named entity recognition techniques. Since

[1]International Union of Pure and Applied Chemistry
[2]International Chemical Identifier

[3]`http://www.drugbank.ca/`, last accessed October 2009

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Patent Retrieval in Chemistry based on Semantically Tagged Named Entities	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Fraunhofer Institute for Algorithms and Scientific Computing (SCAI),Schloss Birlinghoven,53754 Sankt Augustin, Germany,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009) held in Gaithersburg, Maryland, November 17-20, 2009. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**9**	

the entities in chemistry patent space are not as well explored as in biomedical space, we propose to tag the noun phrases and normalize them to their base form before further assessments. From the querying and retrieval point of view, the performance of retrieval using tokens, noun phrases, entities as well as their combinations has been evaluated.

The Sections 2 and 3 describe the workflow used for the technology survey and the prior art search task respectively. Section 4 provides the experimental results of both the tasks and Section 5 provides the concluding remarks.

2 Technology Survey Task

The data provided for the Technology Survey (TS) task contain approximately 1.2 million patents from European and US patent offices, 51,000 full text articles from Royal Society of Chemistry (RSC) and 18 topics that are formulated by human experts as a natural language narrative. The task is to retrieve a set of documents from the corpus that can best answer the question. An example of such as topic is *"TS-7: Please identify documents with formulations of minitabs, containing a Factor Xa inhibitor"*.

2.1 Data Preprocessing

The TREC corpus collection was provided in Extensible Markup Language (XML). As a preliminary measure, an analysis of different sections/zones within the patents and RSC articles was performed. Patent documents contain several fields that are presumably not necessary during retrieval and generate substantial noise while processing the documents. Examples of such fields are *country*, *bibliographic data*, *legal-status*, or *non-English abstracts*. Similar examples within RSC articles are *number of pages*, *citations*, or *editor*. The aim was to use only those fields that have high text/noise ratio and that encompass rich information content. Therefore, with a retrieval point of view, the following fields were chosen to be used for indexing and further assessments:

Patents UCID[4], Publication date, Authors, IPC[5] class, Title, Abstract, Description and Claims.

RSC articles DOI[6], Publication date, Authors, Article body (front) and Article body (back).

2.2 Indexing

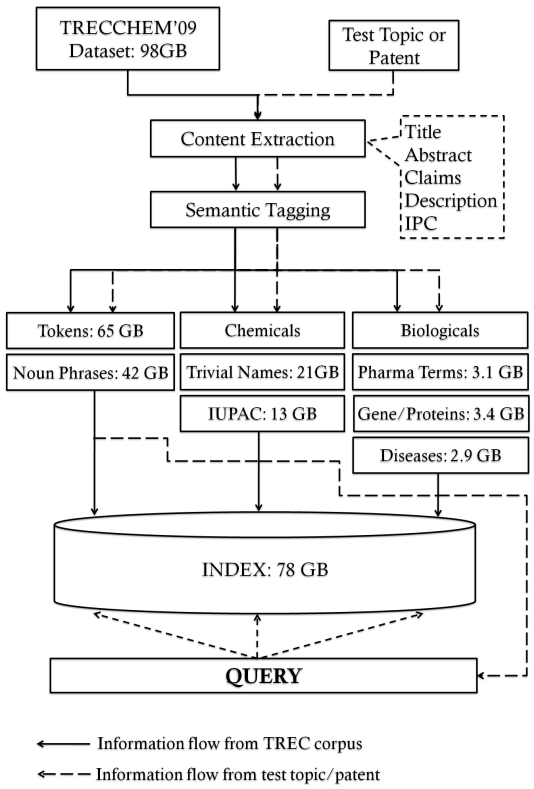

Figure 1: An overview of the workflow implemented for the technology survey and the prior art search tasks. For the technology survey, entity classes occurring within an expert-defined topic (i.e. test topic) are used for querying. Whereas for the prior art search task, entity classes occurring within the test patent are used for querying

[4]User Reference Identifier
[5]International Patent Classification
[6]Digital Object Identifier

Based on preprocessing, the documents were indexed using SCAIView [4]. Figure 1 provides an overview of the methodology used for indexing. SCAIView [4] is a high performing and scalable Information Retrieval (IR) system based on Lucene [3]. It provides a framework for indexing several gigabytes of document data and to quickly perform complex searches over text as well as named entities. The scoring algorithm considers the frequency of a particular term within individual documents and the frequency of the term in all documents. Only the fields mentioned in the previous section were used for indexing while the remaining information were not considered for both patents as well as RSC articles.

2.3 Querying and Retrieval

Considering the time constraints, named entities or noun phrases were not incorporated into the index for this task. Nevertheless, the impact of available semantic dictionaries was tested. The dictionaries used for this task are mentioned in Section 3.2. Three runs have been submitted for the technology survey. For the first two runs, 18 independent topic specific dictionaries were generated whereas for the third run 18 independent noun phrase based queries were used. During the first run, precompiled dictionaries were used for named entity recognition and the results were ordered based on hit frequencies. In the second run, the dictionary entries were queried using SCAIView and the results were ranked based on Lucene's similarity score. For the final run, noun phrases were used for querying and the results were ranked according to the Lucene's similarity score.

2.3.1 Run1: SCAI09TSPM

For this task, the basic idea was to apply named entity recognition and automatically build 18 independent and topic specific dictionaries with the found entities. If the named entities were automatically recognized within TS topic, they were directly used along with their synonyms as present within the dictionary. But as assumed before the dictionaries were not comprehensive to include entities present in all the provided topics. We found no hits in a number of topics, for example in *Synthetic routes used to perform Diels-Alder reaction on a multi-gram scale.* If no hits were detected, the entities were recognized manually and expanded with their synonyms. Example for automatically found and manually generated entries are given in Table 1.

The dictionaries generated for the TS tasks were used within the ProMiner framework [2] for identification of potential term mentions within the corpus of patents as well as RSC articles. The aim was to identify those documents that contain terms present in the dictionary and the documents were ranked based on simple term frequencies.

According to the definition of the TREC task, we were supposed to submit the patent identifiers without 'patent-type' information. Therefore, from all revisions of a patent, the one with maximum score was reported.

2.3.2 Run2: SCAI09TSMAN

This semi-automatic process is intended to give a baseline for retrieval performance of non-patent experts. For this task, the same dictionaries generated for SCAI09TSPM were used. The queries were performed using the SCAIView search engine and the documents were retrieved and ranked based on Lucene's similarity score. The results were filtered to exclude information about the patent-type from the retrieved patent identifiers similar to the process explained in the previous section. Considering the limited set of TS topics, we did not index the chemicals or biomedical terms but rather expanded the queries manually with their corresponding synonyms. An example query for TS-15 is:

```
("Betaine" OR "Glycine betaine" OR
"Glycocol betaine" OR "Glycylbetaine" OR
...)  AND ("Peripheral Artery disorder"
OR "Peripheral Arterial Disease" OR ...)
```

2.3.3 Run3: SCAI09TSNP

The principle behind SCAI09TSNP run was to perform the task in an automatic way based on noun-phrase detection incorporating the OpenNLP chun-

Informative Term	Synonyms	Source
Betaine	Glycine betaine, Glycocol betaine, Glycylbetaine etc.	ATC
Peripheral Artery Disease	Peripheral Artery Disorder, Peripheral Arterial Disease etc.	MeSH
Diels-Alder reaction	Diels Alder reaction Diels Alder mechanism etc.	Manual

Table 1: Synonyms and their sources for informative terms within TS topic

ker[7]. Since noun phrases provide substantial information like head nouns and their modifiers, the idea is to use noun phrases as queries. Noun phrases from each topic description were collected separately and directly used for querying. For this run, the queries were not expanded with synonyms or normalized to unique base forms. An example of query for TS-15 generated with NP chunker is:

```
("cardiovascular" AND "betaines" AND
"peripheral arterial disease")
```

3 Prior Art Search Task

The data provided for the Prior Art (PA) search task contains approximately 1.2 million patents from European and US patent offices and 1000 test/query patents. The task is to retrieve sets of documents from corpus invalidating each query document. An example of such a task is *"PA-1: Find all patents in the collection that would potentially be able to invalidate patent EP-0327505"*.

3.1 Data Preprocessing

The same preprocessing as for the TS task was incorporated for the PA task such as selection of informative fields and extraction of plain text. An analysis of IPC classes was conducted for the large patent corpus as well as query patent subcorpus. The results of IPC class analysis indicate that the superclasses A61 "Medical or Veterinary Science" and C07 "Organic Chemistry" dominate the corpus with more than 70% of the total patents provided.

3.2 Named Entity Recognition

The analysis of IPC classes mentioned in Section 3.1 has shown that organic chemistry, biomedicine and biochemistry occupies a large part of the corpus. The hypothesis is that named entity recognition of chemicals and biomedical terms helps to overcome the problems associated with synonyms by automatic query expansion. ProMiner was used for the task of entity recognition with different dictionaries. Named Entity Recognition was performed independently on Title, Abstract, Claims and Description and indexed separately. The following entity classes have been used:

Chemical Names Chemical names including synonyms, formulae, IUPAC, and brand names of chemical compounds as extracted from DrugBank, KEGG[8] Drugs and KEGG Chemicals. Additionally, **IUPAC-like** names as detected with a machine learning based system [6] are incorporated. It performs an internal normalization to map different variants to one base form.

Genes/Proteins Human genes and protein names as well as their synonyms that are extracted from EntrezGene[9] and UniProt[10] [2].

[7] `http://opennlp.sourceforge.net/projects.html`, last accessed October 2009

[8] `http://www.genome.jp/kegg/`, last accessed October 2009

[9] `http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene`, last accessed October 2009

[10] `http://www.uniprot.org/`, last accessed October 2009

Diseases Disease names and their synonyms that are extracted from the Medical Subject Headings (MeSH)[11].

Pharma Terms Pharmacological terms that are extracted from the Anatomical Therapeutic Chemical (ATC)[12] drug classification system. Since ATC does not contain synonyms and term variants, this information was gathered from UMLS with the help of the MetaMap program [9].

3.3 Noun Phrase Recognition with NP Chunker

As described in Section 2.3, noun phrases designate a good source of information content from text. Therefore, the OpenNLP-based NP chunker was applied to recognize all noun phrases that occur in the query patent corpus that resulted in 549,921 phrases. From the extracted noun phrases, 1000 of them were randomly selected and manually classified as informative or not. Table 2 shows some noun phrases examples for both classes. Since only 52% of the extracted noun phrases were found to be informative, a rule based filtering step was incorporated to remove the non-informative noun phrases. After filtering, 194,322 noun phrases remained with 70% informative terms. In a last step, the noun phrases were normalized using the Norm program[13] provided within Specialist NLP package by National Library of Medicine (NLM). Norm creates an abstract representation of text strings ignoring alphabetic case, inflection, spelling variants, punctuation, genitive markers, stop words, diacritics, symbols, ligatures, and word order. After normalization, the noun phrases with similar base forms were mapped onto each other to generate a noun phrase dictionary which was then used within ProMiner for recognition of potentially useful noun phrases occurring in the patent corpus.

[11] `http://www.nlm.nih.gov/mesh/`, last accessed October 2009

[12] `http://www.genome.jp/kegg/brite.html`, last accessed October 2009

[13] `http://lexsrv3.nlm.nih.gov/SPECIALIST/index.html` , last accessed October 2009

3.4 Indexing

Following data preprocessing and name entity recognition, the document texts as well as the biomedical and chemical entities occurring within them were indexed with SCAIView. Figure 1 shows an overview of the workflow implemented for the PA task.

The entity recognition with different dictionaries was performed separately over title, abstract, description and claim section of each document. Additionally, the entities that occur within different sections were indexed as separate fields. Unlike a conventional index that contains only tokens, the used index additionally contains noun phrases, chemicals and biomedical entities. Finally, the index had 34 fields: ID, Authors, Title, Publication date, IPC class, Abstract, Claims, Description, Chemical names (occurring in Title, Abstract, Claims and Description), IUPAC-like (occurring in Title, Abstract, Claims and Description), Genes/Proteins (occurring in Title, Abstract, Claims and Description), Pharma terms (occurring in Title, Abstract, Claims and Description), Diseases (occurring in Title, Abstract, Claims and Description), and Noun Phrases (occurring in Title, Abstract, Claims and Description). Table 3 shows the frequency of different entities occurring in the entire corpus as well as number of documents that contain at least one entity of interest. Chemical and biomedical entities that are present in our dictionaries occur within a large portion of patent corpus. Noun Phrases do not occur in all 1.2 million patents because only the noun phrases that occur within query corpus were tagged and the remaining noun phrases were excluded.

3.5 Querying and Retrieval

Altogether, 36 runs were submitted for the prior art search task. The queries were performed using different combinations of entity types occurring in the query patent and different search strategies. The results were filtered based on priority dates which means the priority date of the retrieved patent needed to be earlier than the priority date of query patent. In order to understand the importance of filtering, both filtered as well as unfiltered results were sub-

Informative Noun Phrases	Non-informative Noun Phrases
copper strip test	1 2 3 1 2 m 4 R=H
methoxypropynyl group	1200 W 13.56 MHz RF power
biodegradable collagen	about 1800 mg/kg
self-adhesive CODAL tape	A)1>[M M]/(4 [M M] [M M])
tyrosine kinase inhibitor	such difficulties

Table 2: Examples of extracted noun phrases.

	No. of unique entities		No. of documents with one or more entities	
Entity Class	Large Corpus	Query Corpus	Large Corpus	Query Corpus
Chemical Names	14,342	2,661	933,468	869
IUPAC-like	8,504,912	17,972	817,606	629
Pharma Terms	449	193	892,736	431
Genes/Proteins	4,246	639	548,428	246
Diseases	18,458	425	824,415	196
Noun Phrases	182,388	190,528	1,176,217	1000

Table 3: Frequencies of dictionary entries occurring within the the large corpus as well as the query corpus and numbers of documents containing at least one entity of interest.

mitted. Table 5 shows all the submitted runs along with run identifiers, entity types and sections used for querying, and an indication whether the results were filtered or not.

Different objects that were used for querying are:

Tokens Search with all tokens that occur in a query patent

Noun Phrases Search with all noun phrases that occur in a query patent

Entities Search with all chemical entities (chemical names and IUPAC-like) and biomedical entities (pharma terms, genes/proteins and diseases) that occur in a query patent.

The different search strategies are:

Full Document Search in *title*, *abstract*, *claims* and *description*.

Weighted Zones Search different sections of document with different boosting factor. The boosting factors were set to 3 for *abstract*, to 2 for *claims*, to 1.5 for *description* and to 2 for *title*.

Description Only Search within *description* section only.

Claims Only Search within *claims* section only.

Full Document and IPC Class Search in *title*, *abstract*, *claims*, *description* and give high priority to retrieved documents that have same IPC class as query document.

In Table 5, *boosting* indicates a run with a high boosting factor of 3 assigned for all chemical entities, noun phrases or noun phrases that co-occur within same sentences of the query document, respectively. The assumption behind the latter was that co-occurring noun phrases would be descriptive to understand the context of the document and they would serve as a good source for information retrieval.

Run ID	SCAI09TSPM	SCAI09TSMAN	SCAI09TSNP
nDCG	0.357	0.493	0.446

Table 4: Results of the Technology Survey Task. Evaluations are based on nDCG score.

4 Results

4.1 Results of the Technology Survey Task

For the TS task, the reported results are based on the *normalized Discounted Cumulative Gain (nDCG) score* [12]. Table 4 shows *nDCG* scores of all the officially submitted runs for this task.

The run SCAI09TSMAN based on manually formulated queries resulted in the best *nDCG* score of 0.493. The automatic run SCAI09TSNP using only noun phrases without query expansion resulted in a slightly lower score of 0.446. Run SCAI09TSPM using entity recognition and term frequency-based ranking performed worse. This indicates the importance and role of scoring functions as used by Lucene for ranking the relevance of retrieved documents.

4.2 Results of the Prior Art Search Task

For the PA task, the reported results are based on the *Binary Preference score* (*bpref*) [1]. Table 5 shows *bpref* scores for all the officially submitted runs for this task.

The token-based full document search with IPC class outperformed entity-based and noun phrase-based searches. Filtering the results based on the priority date showed mild improvement in the performance of retrieval when compared to unfiltered results. Weighted zone search by boosting different subsections of the document showed to be promising in comparison to normal full document search or only description or claim search. An interesting observation is that the claim search which is broadly employed by patent experts for invalidity search or prior art search reported poor results with token based, noun phrase based as well as entity based search. Weighting/Boosting the entities does not seem to be helpful but a run where only noun phrases were boosted (SCAI09PAf4b) performed slightly better than weighting the entities. An assumption of boosting co-occurring noun phrases (SCAI09PAf4c) during querying indicated a downfall. Nevertheless, the importance of zone weighting and IPC class for patent retrieval was demonstrated.

4.3 Post-TREC Results of the Prior Art Search Task

The analysis of the results of PA task officially submitted to TREC showed that inclusion of zone weighting and the IPC class significantly improves the performance of retrieval. Therefore, utilizing this information, additional experiments were performed with a different combination of entity types used for querying. Table 6 shows the *bpref* scores of post-TREC runs. A combination of tokens, noun phrases and entities searched with zone weighting and IPC class information improved the performance of retrieval. The performance after coupling tokens with noun phrases and tokens with entities showed notable gain in the results. This indicates the essence of using entities and noun phrases for document retrieval as well as combining them in different ways.

5 Discussion and Conclusion

After analyzing the scores achieved during the TS task, the baseline method with manual query formulation and query expansion showed better performance in comparison to the other runs. However, the automatic noun phrase recognition combined with Lucene-based retrieval seems to work considerably well. A better retrieval performance could be achieved through the usage of informative terms and dictionary expansion. For the PA task, the retrieval performance using different entity types as

	Full Document	Weighted Zones	Description Only Only	Claims	Full Document & IPC Class
Tokens	0.3601^{t1a} 0.3777^{f1a}	0.3826^{t1b} 0.3894^{f1b}	0.3336^{t1c} 0.3501^{f1c}	0.2138^{t1d} 0.2137^{f1d}	0.3777^{t1e} $\mathbf{0.4004}^{f1e}$
Noun Phrases	0.3355^{t2a} 0.3418^{f2a}	0.3314^{t2b} 0.3344^{f2b}	0.3405^{t2c} 0.3500^{f2c}	0.2048^{t2d} 0.1990^{f2d}	0.3775^{t2e} 0.3925^{f2e}
Noun Phrases & Entities	0.3369^{t3a} 0.3380^{f3a}	0.3514^{t3b} 0.3536^{f3b}	0.3290^{t3c} 0.3367^{f3c}	0.2105^{t3d} 0.2035^{f3d}	0.3726^{t3e} 0.3811^{f3e}
Noun Phrases & Entities (Boost Chemicals)	0.3166^{t4a} 0.3181^{f4a}	N/A N/A	N/A N/A	N/A N/A	N/A N/A
Noun Phrases & Entities (Boost Noun Phrases)	0.3666^{t4b} 0.3734^{f4b}	N/A N/A	N/A N/A	N/A N/A	N/A N/A
Noun Phrases & Entities (Boost co-occurring NP)	0.3440^{t4c} 0.3485^{f4c}	N/A N/A	N/A N/A	N/A N/A	N/A N/A

Table 5: Results of the Prior Art Search Task. Scores having 't' and 'f' within run identifier indicates that the results were unfiltered and filtered respectively. The last three characters of the run identifiers are mentioned in the table. An example of submitted run identifier looks like 'SCAI09PAt1a'. The entity types and document sections used for querying are also mentioned. Evaluations are based on *bpref* score.

Tokens & NP & Entities	Tokens & NP	Tokens & Entities
0.4355	0.4302	0.4121

Table 6: Results of the post-TREC Prior Art Search Task. Results are filtered and the evaluations are based on *bpref* score

well as different search strategies was demonstrated along with the importance of zone weighting and using meta-information like IPC class for patent retrieval. Finally, it was shown that querying with a combination of tokens, noun phrases and entities performed relatively better than using the different entity types alone.

There are several ways to improve the performance of retrieval. Currently, the breadth of knowledge sources that has been used is limited. For example, only the chemicals present in DrugBank and KEGG databases have been used. These databases are specialized to include compounds that are of biomedical interest and does not focus on chemicals contained in ink formulations, cement or fertilizers. Considering the scope of IPC classes of the documents provided within the TREC data set, only 50% of the documents belong to the biomedical domain. Therefore, indexing the terms using broader resources that cover terminologies beyond the biomedical domain has to be tested in future approaches. Using a pre-trained NP chunker [5] that has been specifically trained on chemistry-based patents is one way to reduce the noisy noun phrases. Improving the recognition performance of the entity recognizers on patents can also contribute to better retrieval. The methods pre-

sented here adopt most of the strategies from conventional document retrieval techniques. However, being at an early stage of patent retrieval, the circumstances underpin the necessity for methods specialized for patent text analysis. Our future work will focus on overcoming the limitations that have been mentioned previously and to optimize our retrieval system to better adapt to chemistry-based patents.

References

[1] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, 2004.

[2] Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 Suppl 1:S14, 2005.

[3] Erik Hatcher and Otis Gospodnetic. *Lucene in Action (In Action series)*. Manning Publications Co., Greenwich, CT, USA, 1st edition, 2004.

[4] Martin Hofmann-Apitius, Juliane Fluck, Laura Furlong, Oriol Fornes, Corinna Kolářik, Susanne Hanser, Martin Boeker, Stefan Schulz, Ferran Sanz, Roman Klinger, Theo Mevissen, Tobias Gattermayer, Baldo Oliva, and Christoph M Friedrich. Knowledge environments representing molecular entities for the virtual physiological human. *Philos Transact A Math Phys Eng Sci*, 366(1878):3091–3110, Sep 2008.

[5] Joachim Wermter, Juliane Fluck, Jannik Strötgen, Stefan Geißler, and Udo Hahn. Recognizing Noun Phrases in Biomedical Text: An Evaluation of Lab Prototypes and Commercial Chunkers. In *SMBM 2005 - Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine*, pages 25–33, 2005.

[6] Roman Klinger, Corinna Kolářik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. Detection of IUPAC and IUPAC-like Chemical Names. *Bioinformatics*, 24(13):i268–i276, 2008. Proceedings of the International Conference Intelligent Systems for Molecular Biology (ISMB).

[7] Corinna Kolářik, Roman Klinger, Christoph M. Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. Chemical Names: Terminological Resources and Corpora Annotation. In *Workshop on Building and evaluating resources for biomedical text mining*, volume 6th edition of the Language Resources and Evaluation Conference, pages 51–58, Marrakech, Morocco, 2008.

[8] A. D. McNaught and A. Wilkinson. *Compendium of Chemical Terminology – The Gold Book*. Blackwell Science, 1997.

[9] John D Osborne, Simon Lin, Lihua Zhu, and Warren A Kibbe. Mining biomedical data using MetaMap Transfer (MMTx) and the Unified Medical Language System (UMLS). *Methods Mol Biol*, 408:153–169, 2007.

[10] Dolf Trieschnigg, Wessel Kraaij, and Martijn Schuemie. Concept based document retrieval for genomics literature. In *TREC Genomics Track*, pages 1–11, 2006.

[11] Jay Urbain, Nazli Goharian, and Ophir Frieder. IIT TREC-2007 Genomics Track: Using Concept-based Semantics in Context for Genomics Literature Passage Retrieval. In *TREC Genomics Track*, pages 1–4, 2007.

[12] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610, 2008.